# **EXPERIMENT REPORT Part B**

| **Student Name** | Shalimar Chalhoub |
| --- | --- |
| **Project Name** | Regression Models |
| **Date** | 31/3/2023 |
| **Deliverables** | <MLAA Assignment 1 Part B>  <Multivariate Linear Regression> |

| 1. **EXPERIMENT BACKGROUND** | | |
| --- | --- | --- |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | | |
| **1.a. Business Objective** | The Business objective behind this project is to accurately predict cancer mortality based on information related to US counties By doing a multivariate Linear Regression.  By interpreting the results, the business can be able to know which factors affect the mortality rate due to cancer and can work towards lowering said rate.  Incorrect results could lead the business to overlooking some counties that need urgent help in lowering the mortality rate as the result was inaccurate | |
| **1.b. Hypothesis** | Are we able to predict the deathRate due to cancer of each US county using multiple features from the dataset without doing feature engineering?  This model is worthwhile considering, because it takes into account all of the variables in the dataset and uses all of them to predict the target rate, which in theory, should perform good. | |
| **1.c. Experiment Objective** | I am expecting to build a decent Linear Regression model that will accurately predict the deathRate in each county with minimal error.  I am aware from my given experiment that the data is not very linear, thus I am expecting a lower MSE than before.  As a result, it could either predict a good model which we will be able to use, or a model that presents a lot of errors and thus, is unusable. | |

| 1. **EXPERIMENT DETAILS** | | |
| --- | --- | --- |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | | |
| **2.a. Data Preparation** | The below steps were taken for Data Preparation   1. Only took deathRate values between 130 and 240 2. Dropped all rows with an avgAnnCount of 1962.6676684 because, by looking at the dataset, it looked like an outlier 3. Removed all non numeric columns 4. Removed PctSomeCol18\_24 from the dataset because it had more than three fourth of its data missing 5. Removed columns with low correlation such as PctBlack, PctAsian and PctWhite as well as Id column 6. Replaced PctEmployed16\_Over missing values with the mean 7. Split the train set to dev and train (20-80) | |
| **2.b. Feature Engineering** |  | |
| **2.c. Modelling** | The model used for this report is a multivariate linear which is used to model the relationship between multiple independent variables and a single dependent variable.  I used all the numeric functions in the dataset with the exception of PctSomeCol18\_24,PctBlack, PctAsian, PctWhite and Id due to irrelevancy.  was chosen because I have been imposed the constraint of having to use it. | |

| 1. **EXPERIMENT RESULTS** | | |
| --- | --- | --- |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | | |
| **3.a. Technical Performance** | This model performed better than the univariate Linear regression and produced an MSE of 306.901 which means that there are 17.51 data points between real values and predicted values and whilst this is less than the univariate linear regression, it still is not a good result.  The cause of that error is that the data is not linear and do not have a linear relationship | |
| **3.b. Business Impact** | This experiment does not meet the business objectives as there is a big error margin and hence, will not predict accurate results.  Counties with high DeathRate which have been predicted as low, will be overlooked by the business which might result in higher death rates.  Also, counties with low DeathRates are predicted as high, resulting in the company investing more money and resources looking into why, when in reality, there is no need for it | |
| **3.c. Encountered Issues** | During this project I encountered issues with outliers, which I solved by setting bounds to my data. I also had some missing values that I fixed as well as some irrelevant features that made my MSE higher so I removed those.  Some issues that I couldn’t fix is the non-linearity of the data, for which I have no solution due to he restrictions on this part of the assignment | |

| 1. **FUTURE EXPERIMENT** | | |
| --- | --- | --- |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | | |
| **4.a. Key Learning** | Using all of the features definitely produced a better outcome than using just one and this model performed better than the univariate Linear regression. Removing some columns really helped lowering the error margin and getting better results.  It can be seen that better results are yet to be produced and since the data is non linear, we can do some more data transformation in order to be able to get better results, thus, this is not a dead end and without restrictions, better results can be achieved. | |
| **4.b. Suggestions / Recommendations** | After analyzing the results of the model, we can see that better results can be achieved, and so for the next step, we have to look at feature engineering whilst trying to make the data more linear and perform another kind of Regression in order to achieve better results.  This model can be tested by getting data from years prior and trying the model on them in order to see if the predictions are constant.  Upon receiving the desired results, it can be used to predict deathRates and work towards lowering them by either investing in research or better healthcare | |